STAT 547M, Basic Training in Data Science
2017/18, Term 1
Instructors: Vincenzo Coia and Giulio Dalla Riva

Time and Place: T/R, 9:30-11:00 AM, Oct 24,- Dec 01, ESB 2012

Course webpage: http://stat545.com

Objectives: Students will learn
-to explore, groom, visualize and analyze data
-make all of that reproducible, reusable, and shareable
-using R

Prerequisites: Stat 545A. Please consult the instructor if in doubt.

Textbook/course materials: All course material is presented in the class or online

Assessment: Course mark = 75% homework, 15% peer review, 10% engagement
(i.e. attendance, participation in discussion in-class and/or online)

Students submit homework once a week. The end result may still resemble a "final project", but one that has been built up gradually over the term.

Topics: Selected topics may include (but are not limited to):

- Introduction to <u>R</u> and the <u>RStudio IDE</u>; scripts, the workspace, RStudio Projects
- Generate reports from R scripts and <u>R Markdown</u>
- Care and feeding of data in R
- Data aggregation; "apply" functions, `plyr`, `dplyr`
- Data visualization with `ggplot2`
- Graphs and descriptive stats for quantitative and categorical variables
- Writing R functions
- Coding style and project organization
- Version control with Git; collaboration via <u>GitHub</u>
- Character data; regular expressions
- Interactive pages, apps, and graphics with <u>Shiny</u> and `ggvis`
- Get data off the web and expose data, code, results on the web
- Distribute data and code via an R package
- Automate an analytical pipeline, e.g. via `Make`

Consult the course webpage for more information on specific topics to be covered.